Published online 18 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D957–D963***
doi:10.1093/nar/gkr959

COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines

Judice L. Y. Koh[1], Kevin R. Brown[1], Azin Sayad[1], Dahlia Kasimer[1], Troy Ketela[2] and Jason Moffat[1,*]

[1]Banting and Best Department of Medical Research, Department of Molecular Genetics, Donnelly Centre, University of Toronto, 160 College St., Toronto, ON M5S 3E1 and [2]Ontario Institute for Cancer Research, 101 College St Suite 800, Toronto, ON, Canada M5G 0A3

Received August 19, 2011; Revised September 6, 2011; Accepted October 10, 2011

ABSTRACT

Genome-wide pooled shRNA screens enable global identification of the genes essential for cancer cell survival and proliferation and provide a 'functional genetic' map of human cancer to complement genomic studies. Using a lentiviral shRNA library targeting approximately 16 000 human genes and a newly developed scoring approach, we identified essential gene profiles in more than 70 breast, pancreatic and ovarian cancer cell lines. We developed a web-accessible database system for capturing information from each step in our standardized screening pipeline and a gene-centric search tool for exploring shRNA activities within a given cell line or across multiple cell lines. The database consists of a laboratory information and management system for tracking each step of a pooled shRNA screen as well as a web interface for querying and visualization of shRNA and gene-level performance across multiple cancer cell lines. COLT-Cancer Version 1.0 is currently accessible at http://colt.ccbr.utoronto.ca/cancer.

BACKGROUND

Over the past decade, gene silencing through RNA interference (RNAi) technology has emerged as a powerful tool for deciphering the mechanistic details of biological processes in higher eukaryotes. RNAi was first exploited for systematic functional studies in *Caenorhabditis elegans* and *Drosophila melanogaster* (1,2), and is now also widely used for selective suppression of gene expression in mammalian cells (3,4). More recently, viral-based pooled shRNA screening methods have been developed and applied in functional genetic screens to identify genes that are essential for cancer cell proliferation with the goal of identifying therapeutic targets (5–7).

We have developed a standard operating procedure for carrying out large-scale pooled shRNA screens and are systematically looking for essential genes using cancer cell lines from various tumor types including breast, ovarian and pancreatic (Marcotte *et al.*, submitted for publication). We are using a pooled subset of the human TRC collection that includes 78 432 shRNAs targeting approximately 16 000 human genes (5,8) to develop essential gene profiles across a large number of cancer cell lines (R. Marcotte, submitted for publication).

To streamline the process from data generation to public user access of our cancer cell line screening results, we developed a web-accessible database system for processing, analyzing and retrieving data from the pooled screens. The COLT-Cancer database system is comprised of: (i) a laboratory information and management system (LIMS) for automation of basic microarray functions such as chip signal extraction, background correction, normalization and quality metric generation; (ii) an automated routine for generating hairpin-level and gene-level essentiality scores; and (iii) a web interface at http://colt.ccbr.utoronto.ca/cancer that enables researchers to query, visualize and compare essential genes across multiple cancer cell lines.

Many RNAi screens in mammalian cells have been conducted in academic and industry labs and have yielded novel insight into genes that are essential for cancer cell proliferation. Some of the resulting data is available to the research community through a number of collation efforts, including RNAiDB (9), GenomeRNAi (10) and FLIGHT database (11). These databases support integrative visualization and analysis of RNAi data with other data such as gene annotations, shRNA sequence annotations and corresponding knockout efficiency and genomic information. In addition, several RNAi-based tools/databases also focus on providing searchable shRNA and siRNA constructs, such as RNAi Codex (12), E-RNAi (13), the RNAi Consortium (TRC) library database (http://www.broadinstitute.org/rnai/public/) and the

*To whom correspondence should be addressed. Tel: +1 416 978 0336; Fax: +1 416 978 8287; Email: j.moffat@utoronto.ca

Cancer Genome Anatomy Project (CGAP) shRNA clone library (http://cgap.nci.nih.gov/RNAi/RNAi2). The COLT-Cancer was designed with a unique focus to facilitate functional comparison of essential gene profiles across a compendium of cancer cell lines and integrate this information with structural genomic data from large cancer genome sequencing efforts to uncover vulnerabilities that can be used to develop better prognostics and therapeutics (Marcotte *et al.*, submitted for publication).

DATABASE CONSTRUCTION AND CONTENT

System architecture

COLT-Cancer is deployed on a back-end DB2 relational database management system. The DB2 database serves as central storage for data and images generated continuously from our automated computational pipeline for processing and analyzing RNAi pooled screens. As such, it was designed with the objective of achieving query and storage efficiency for large quantities of microarray images, signal intensity measurements, annotations of genes and shRNA reagents, genomic information, and other metadata and contains more than 200 relational tables (database schema available at COLT-Cancer online documentation). The COLT system is hosted on 2 IBM servers; one that functions as a database server and the other as a web server to facilitate querying, data downloading and data visualization through the COLT-Cancer websites. The web interfaces of COLT-Cancer were developed using a combination of HTML, CGI Perl, DB2/Perl application programming interface, cascading style sheets and Javascript for easy navigation. Graphical plots are generated on-the-fly using R plotting functions.

Microarray LIMS system

At the back-end of COLT is a LIMS to support basic tasks central to our standard operating procedures such as signal extraction, background correction, normalization and quality metric generation (Figure 1a). The data extraction and processing pipeline is fully automated, and is triggered upon receipt of a .CEL file. Raw microarray data is extracted and background corrected using Affymetrix Power Tools (APT) v1.12.0 (http://www.affymetrix.com). There are 33 894 features on the Gene Modulation Array Platform (GMAP) used for background correction (8). Replicate GMAP arrays are normalized using Cyclic Loess based on the 'MA-plot' of pairs of arrays and performed via the 'affy' R package (14). In general, the GCbg-correction increases differentiation between feature signals from the pooled and background probes, while normalization reduces variance between replicates (Figure 1b). To allow access to raw and normalized shRNA, spike-in and control features on all microarrays for quality control checks and analyses beyond the database functions, a web interface was developed to provide authenticated access to internal and external end-users. For published experiments, these are accessible through the 'COLT LIMS' link on COLT-Cancer.

Quantifying drop-out rates of shRNA and target genes

To score screening data, shRNA-level shARP (shRNA Activity Ranking Profile) and gene-level GARP (Gene Activity Rank Profile) scores were computed across time points in each experiment. Details of the scoring methods are described elsewhere (Marcotte *et al.*, submitted for publication; and in online documentation of website). Simply, the average of the two lowest shARP scores for a given gene is used to calculate the GARP score for that gene. The lower the GARP score for a given cell line, the more essential is the gene in that particular cell line. *P*-values are calculated from permutation testing of 1000 random scores, as a measure of the statistical 'confidence' of the GARP score. The GARP, GARP *P*-values, and shARP scores are available for download on COLT-Cancer.

Integration with external data sources

To add contextual information to shRNA screen scoring results, COLT-Cancer provides specific types of annotations for each gene. These annotations include: cancer-associated genes, known activity as a House-keeping (HK) protein, predicted or known localization to the cell surface, associations with NCI curated pathways and copy number abberation (CNA) data from a number of breast and pancreatic tumor samples.

HK genes are vital for maintaining the biological 'well-being' of cells under various conditions, and are therefore ubiquitously expressed across tissues and cell types and are evolutionally more conserved (15,16). Previous studies have shown that HK genes are highly associated with essentiality, and with somatic cancer and other diseases (17,18), thus they can be used to benchmark the essentiality of the gene from our experiments against relevance to HK functions and oncogenesis. The known HK genes in COLT-Cancer were obtained from Tu *et al.* (18).

Not all genes required for survival in cancer cells are directly associated with oncogenesis. On the other hand, comprehensive knowledge of gene essentiality may unravel novel drivers of cell survival. To help dissect known cancer-associated essential genes from essential genes of unknown significance, we integrated gene–cancer associations from the Cancer Gene Consensus (CGC, http://www.sanger.ac.uk/genetics/CGP; 19). The CGC currently catalogs 291 genes that show somatic and/or germline mutations in cancer. Additionally, we integrated curated cancer-associated pathways from the Pathway Interaction Database (PID) (20).

Proteins located on the cell surface are of interest for identifying putative antibody-based tumor targets. Two 'surfaceome' data sets—a comprehensive set of 3702 genes predicted from their transmembrane domain and gene ontology annotations using bioinformatic approaches (21), and a larger curated set of surfaceome genes (R.J. Williams and J. Dennis, personal communication). These data sets are downloaded on regular basis, and reformatted into the data model, in order to speed up web queries.

CNAs are known to be characteristics of certain cancers, and may potentially serve as a benchmark for identifying novel oncogenes from the top-ranked essential genes in COLT-Cancer. All cell line copy number data were provided by the Cancer Genome Project group at

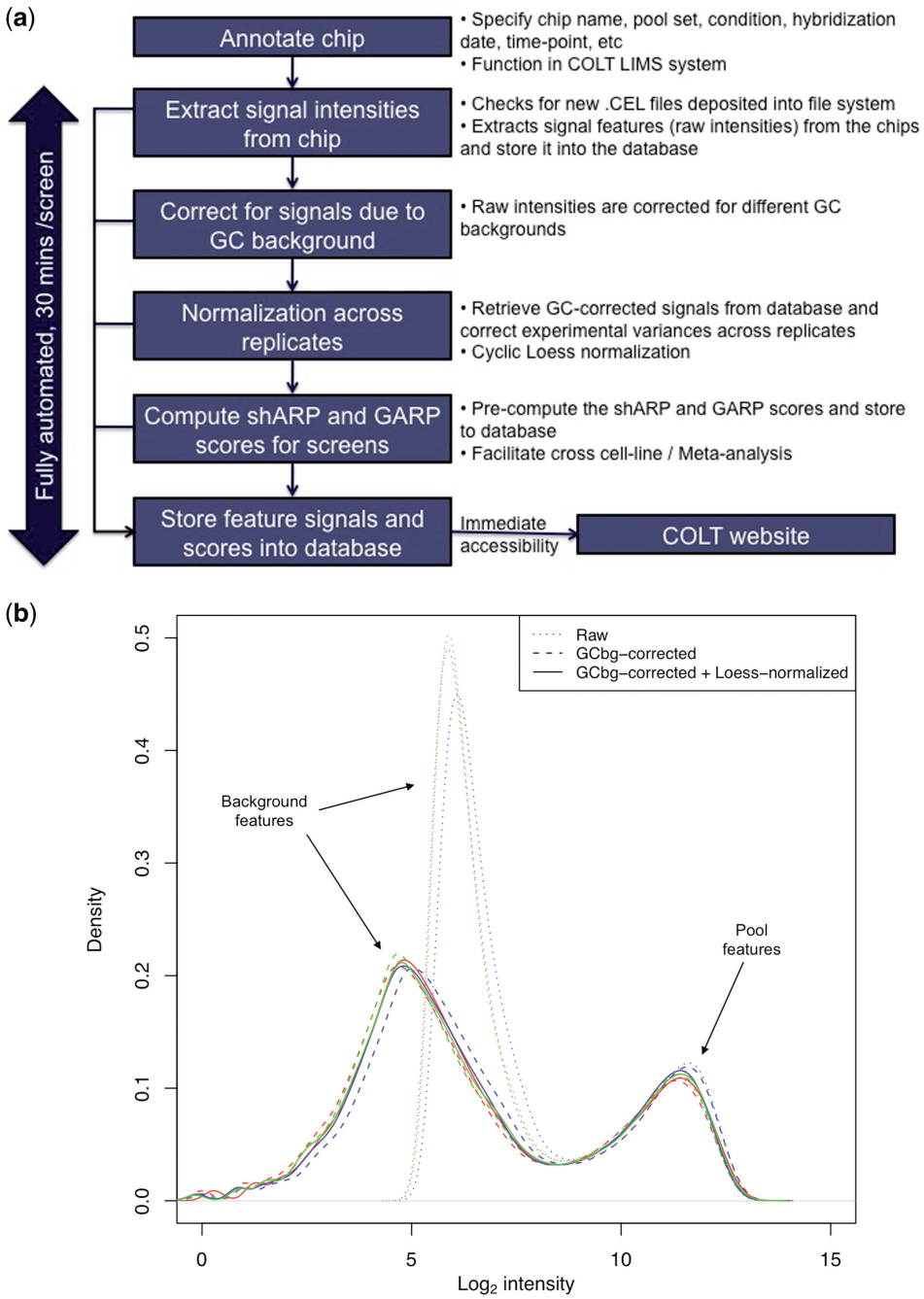

Figure 1. (**a**) Microarray data extraction and processing pipeline. (**b**) Distribution of a sample experiment from COLT-Cancer. GCbg-correction increases differentiation between feature signals from the pooled and background probes, while normalization reduces variance between replicates.

the Wellcome Trust Sanger Institute and can be obtained from http://www.sanger.ac.uk/genetics/CGP/CellLines/. The Cancer Genome Atlas (TCGA, 22) was used for ovarian primary tumor copy number data. The University Hospital Ullevål (Ull) cohort breast samples were used for breast (23). For pancreatic tumor data, we used the data published by Harada *et al.* (24). These samples were processed using the Aroma.CN package (25,26), including Single Nucleotide Polymorphism-Robust Multi-Array Analysis (SNP-RMA, 27) normalization. In all cases, data was segmented using Circular Binary Segmentation (CBS, 28) and a Log-R ratio threshold of 0.33 was used to call copy number gain.

DATABASE UTILITY

Browse cell lines and genes

COLT-Cancer currently hosts 72 published human cancer cell lines. Complete lists of the cell lines and genes are alphabetically ordered by their common names and available to users through the 'List cell-lines' and 'List all

genes' functions. They can also be independently searched using various options including name, reference sequence, gene IDs and gene descriptions. Cell line information also include short-tandem repeat profiles and are linked to ATCC Cell Biology Collection.

Gene-centric queries across cell lines

Users can search for genes-of-interest across selected tumor types or subtypes. A gene-centric search result returns the GARP scores and *P*-values of the best matching gene across selected screens, along with externally integrated annotations of the gene (HK, CGC, surfaceome, among others). End-users can choose to download the scores or evaluate similarities amongst genes or cell lines through hierarchical clustering of either GARP scores or *P*-values (Figure 2).

Cross cell-line comparisons

Web-based functions are provided to facilitate functional comparisons across studies, which are critical for understanding the common and discriminating machineries in different cancer cell lines. For instance, the 'Cross cell-line search' option on COLT-Cancer allows users to query, visualize and compare genes that are commonly essential across multiple tumor types or within the same tumor type in a summarized matrix view (Figure 3).

Essentiality and CNA plot

These query plots provide an integrated representation of essentiality and CNA data for contiguous genes within a genomic region, across three tumor types including breast, ovarian and pancreatic. They allow users to examine the degree of essentiality for a given gene within a specified genomic region, and how the degree of essentiality is distributed across breast, ovarian and pancreatic cancer cell lines. Bars above the axis represent the fraction of all screens in which a given gene had an essentiality score in the top 5%, by GARP. Bars below the axis represent the fraction of samples, derived from literature-curated data, in which the gene is amplified in either cell lines or patient tumors.

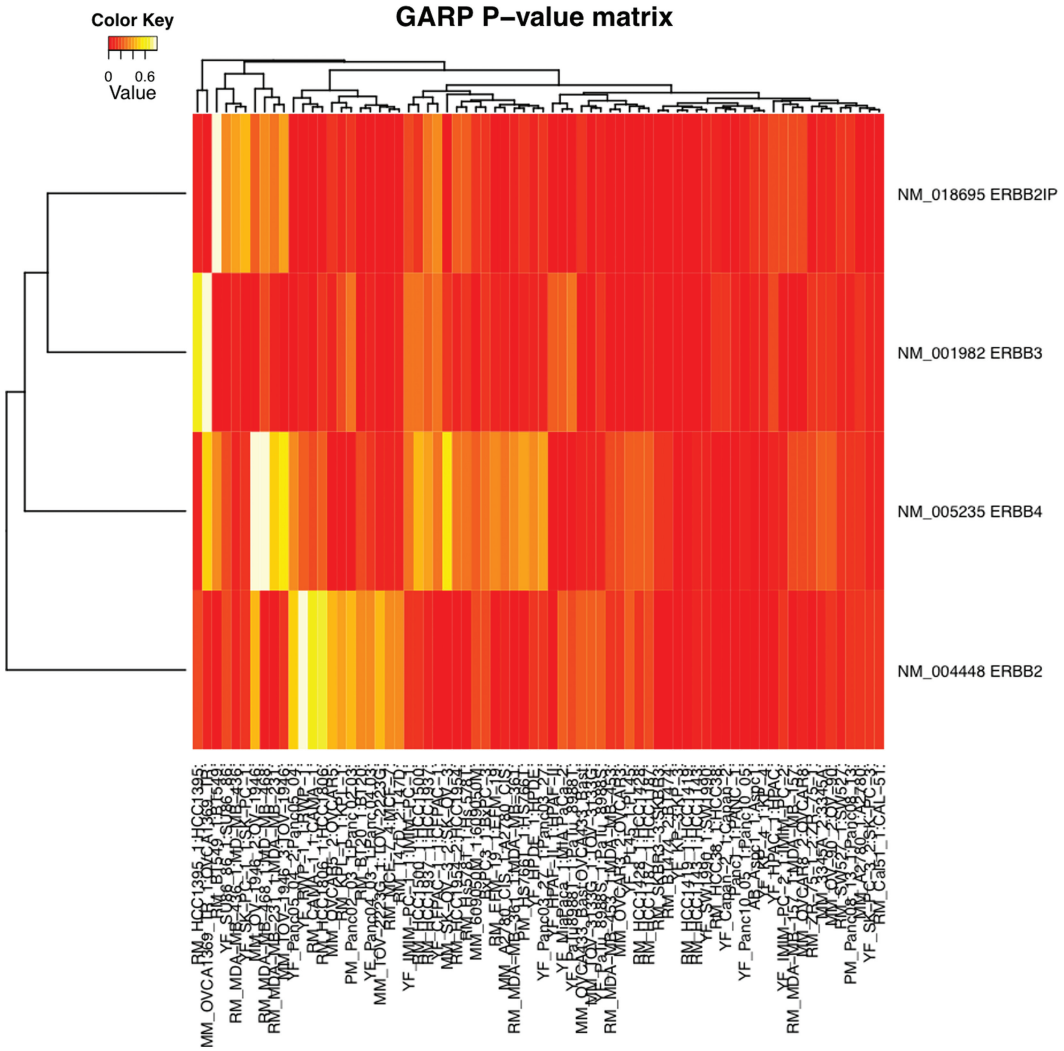

Figure 2. An example of a hierarchically clustered heatmap generated in COLT-Cancer from the GARP P-values of 4 genes across all cancer cell lines.

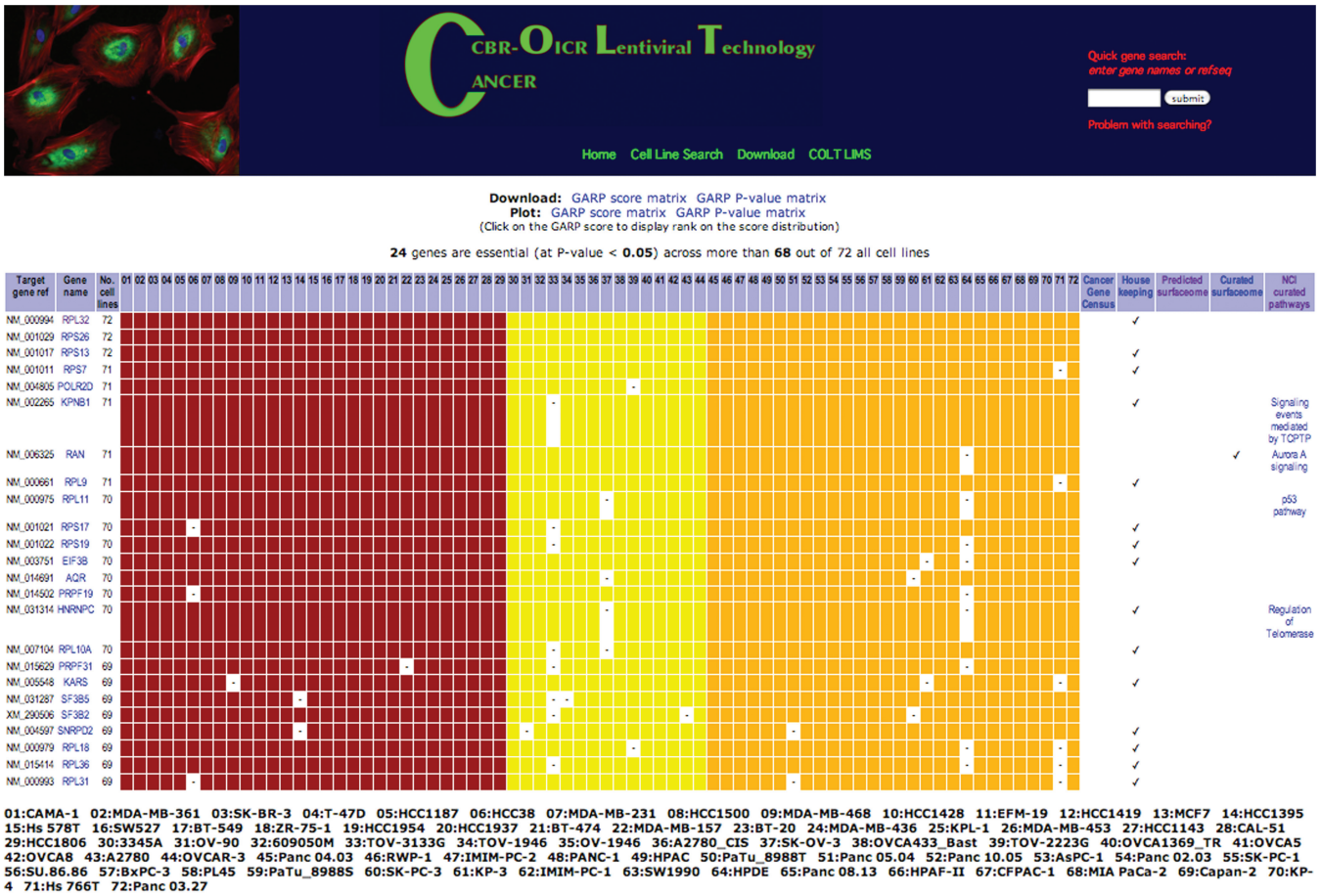

Figure 3. An example use case in COLT-Cancer to visualize common essential genes across breast (red), ovarian (yellow) and pancreatic (orange) cell lines.

COLT LIMS

Signal features, along with descriptions of the gene targeted by the shRNA sequence can be downloaded into tab-delimited files. On-the-fly visualization, basic statistical analysis, and quality metrics are integrated, allowing end-users to assess the quality of the screens. For example, the 'Distribution plot' presents a histogram-like view of all shRNA signals from selected chips in an experiment, allowing the user to assess replicate reproducibility, and provides a visual approximation of the number hairpins dropping out of the screen. The 'Correlation plot' displays the Spearman correlation between all pairs of selected chips (Figure 4), providing another assessment of screen quality and replicate reproducibility.

Download

COLT-Cancer supports downloading of user-specified subsets or the full set of essentiality GARP scores and corresponding *P*-values in tab-delimited files. Microarray images can be downloaded as lower resolution JPEG files.

Future developments

COLT-Cancer was developed with the intent to be an active resource for quantitative gene essentiality measurements of human cancer cells. There are mechanisms for future regular updates of experimental results through the COLT LIMS pipeline. Future additions to the system aim at further automating the updating processes to enable more frequent updates of external data sources.

Although COLT-Cancer currently contains only locally generated pooled shRNA screening data, further work will involve integrating data types from other sources. Future versions of COLT-Cancer also aim to provide new visualization tools such as integrating our results into a genome viewer to enrich COLT-Cancer as a central resource for cancer research, assisting experimentalists in the generation of testable hypotheses.

CONCLUSIONS

We present a web database referred to as COLT-Cancer, which provides users an integrative platform to browse, query and analyze high-quality pooled shRNA screening data from a deep repository of functional genetic screens across different tumor types and genetic backgrounds. The COLT-Cancer website is a valuable resource to identify genetic vulnerabilities for cancer cell proliferation that may serve as useful prognostic or therapeutic targets.

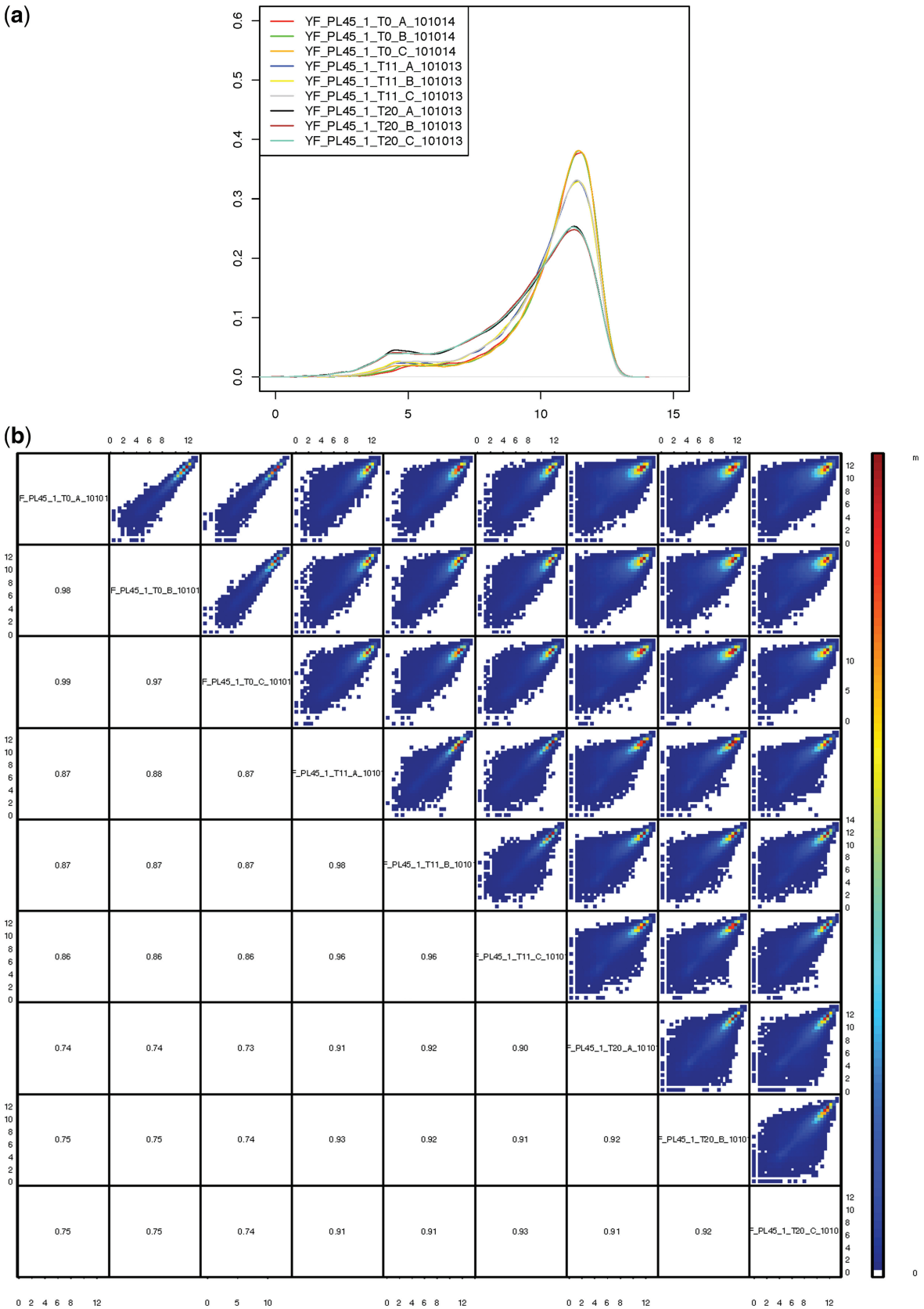

Figure 4. (**a**) A plot generated on-the-fly from COLT-LIMS shows the distribution/histogram of the signals from each chip in a screen. (**b**) Spearman correlations between pairs of microarrays in a sample screen.

ACKNOWLEDGEMENTS

The authors thank Richard Marcotte, Fernando Suarez, Fabrice Sircoulomb, Paul M. Krzyzanowski, Franco Vizeacoumar, Rob Rottapel, Ben Neel and all the members of the Moffat lab and anonymous users of COLT-Cancer beta version for testing and suggestions.

FUNDING

Ontario Institute for Cancer Research and Terry Fox Research Institute through the Selective Therapies Program (J.M.); the Canadian Institutes for Health Research #178975 (J.M.); the Canadian Foundation for Innovation (J.M.); and the Ontario Research Fund (J.M.). J.M. holds a CIHR New Investigator Award. Funding for open access charge: OICR/TFRI Selective Therapies Program.

Conflict of interest statement. None declared.

REFERENCES

1. Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
2. Hammond,S.M., Bernstein,E., Beach,D. and Hannon,G.J. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, **404**, 293–296.
3. Brummelkamp,T.R., Bernards,R. and Agami,R. (2002) A system for stable expression of short interfering RNAs in mammalian cells. *Science*, **296**, 550–553.
4. Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
5. Moffat,J., Grueneberg,D.A., Yang,X., Kim,S.Y., Kloepfer,A.M., Hinkle,G., Piqani,B., Eisenhaure,T.M., Luo,B., Grenier,J.K. *et al.* (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, **124**, 1283–1298.
6. Silva,J.M., Marran,K., Parker,J.S., Silva,J., Golding,M., Schlabach,M.R., Elledge,S.J., Hannon,G.J. and Chang,K. (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, **319**, 617–622.
7. Schlabach,M.R., Luo,J., Solimini,N.L., Hu,G., Xu,Q., Li,M.Z., Zhao,Z., Smogorzewska,A., Sowa,M.E., Ang,X.L. *et al.* (2008) Cancer proliferation gene discovery through functional genomics. *Science*, **319**, 620–624.
8. Ketela,T., Heisler,L.E., Brown,K.R., Ammar,R., Kasimer,D., Surendra,A., Ericson,E., Blakely,K., Karamboulas,D. and Smith,A.M. (2011) A comprehensive platform for highly multiplexed mammalian functional genetic screens. *BMC Genomics*, **12**, 213.
9. Gunsalus,K.C., Yueh,W.C., MacMenamin,P. and Piano,F. (2004) RNAiDB and PhenoBlast: Web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.*, **32**, D406–D410.
10. Horn,T., Arziman,Z., Berger,J. and Boutros,M. (2007) GenomeRNAi: a database for cell-based RNAi phenotypes. *Nucleic Acids Res.*, **35**, D492–D497.
11. Sims,D., Bursteinas,B., Gao,Q., Zvelebil,M. and Baum,B. (2006) FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. *Nucleic Acids Res.*, **34**, D479–D483.
12. Olson,A., Sheth,N., Lee,J.S., Hannon,G. and Sachidanandam,R. (2006) RNAi Codex: a portal/database for short-hairpin RNA (shRNA) gene-silencing constructs. *Nucl. Acids Res.*, **34**, D153–D157.
13. Horn,T. and Boutros,M. (2010) E-RNAi: a web application for the multi-species design of RNAi reagents—2010 update. *Nucleic Acids Res.*, **38**, W332–W339.
14. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
15. Watson,J.D., Hopkins,N.H., Roberts,J.W., Steitz,J.A. and Weiner,A.M. (1965) *Molecular Biology of The Gene*, Vol. 1. Benjamin/Cummings, Menlo Park, California, p. 704.
16. Zhang,L. and Li,W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.*, **21**, 236–239.
17. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabási,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
18. Tu,Z., Wang,L., Xu,M., Zhou,X., Chen,T. and Sun,F. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 3.
19. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
20. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2008) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
21. da Cunha,J.P., Galante,P.A., de Souza,J.E., de Souza,R.F., Carvalho,P.M., Ohara,D.T., Moura,R.P., Oba-Shinja,S.M., Marie,S.K., Silva,W.A. Jr *et al.* (2009) Bioinformatics construction of the human cell surfaceome. *Proc. Natl Acad. Sci. USA*, **106**, 16752–16757.
22. Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
23. Russnes,H.G., Vollan,H.K., Lingjaerde,O.C., Krasnitz,A., Lundin,P., Naume,B., Sørlie,T., Borgen,E., Rye,I.H., Langerød,A. *et al.* (2010) Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl Med.*, **2**, 38–47.
24. Harada,T., Chelala,C., Bhakta,V., Chaplin,T., Caulee,K., Baril,P., Young,B.D. and Lemoine,N.R. (2008) Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene*, **27**, 1951–1960.
25. R Development Core Team, R Foundation for Statistical Computing. (2011) R: A Language and Environment for Statistical Computing.
26. Bengtsson,H., Simpson,K., Bullard,J. and Hansen,K. (2008) aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Technical Report* #745. University of California Department of Statistics, Berkeley.
27. Carvalho,B., Bengtsson,H., Speed,T.P. and Irizarry,R.A. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.
28. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.